第 58 卷 第 24 期/2021 年 12 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

# 基于注意力机制和 Soft-NMS 的改进 Faster R-CNN 目标检测算法

王凤随[1,2,3*]，王启胜[1,2,3]，陈金刚[1,2,3]，刘芙蓉[1,2,3]

[1]安徽工程大学电气工程学院，安徽 芜湖 241000；

[2]检测技术与节能装置安徽省重点实验室，安徽 芜湖 241000；

[3]高端装备先进感知与智能控制教育部重点实验室，安徽 芜湖 241000

**摘要** 针对目标检测网络 Faster R-CNN(Faster Region-Convolutional Neural Network)存在漏检、误检和检测精度低的问题，提出一种融合注意力机制和 Soft-NMS(Soft Non-Maximum Suppression)的 Faster R-CNN 目标检测算法。为了增强 Faster R-CNN 目标检测算法对特征图中全局重要特特的提取并弱化无关特征，首先在网络中引入了注意力机制；其次针对注意力机制采用两个全连接层构成瓶颈结构会造成局部信息损失的问题，构建一种可以和卷积神经网络进行端到端训练的非降维通道注意力和空间注意力串联模块；然后通过在区域建议网络中引入 Soft-NMS 替换传统的非极大抑制算法，可以降低目标漏检并提高定位精度；最后在评价标准里引入了误检率，进一步验证模型的性能。实验结果表明，基于 ResNet-50 的 Faster R-CNN 目标检测算法有效降低了漏检、误检并提高了定位精度，而且在平均检测精度上得到了明显的提升。

**关键词** 光计算；目标检测；注意力机制；非极大抑制；卷积神经网络；Faster R-CNN

**中图分类号** TP181 **文献标志码** A **doi**：10.3788/LOP202158.2420001

# Improved Faster R-CNN Target Detection Algorithm Based on Attention Mechanism and Soft-NMS

Wang Fengsui[1,2,3*], Wang Qisheng[1,2,3], Chen Jingang[1,2,3], Liu Furong[1,2,3]

[1] *School of Electrical Engineering, Anhui Polytechnic University, Wuhu, Anhui* 241000, *China*;

[2] *Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Wuhu, Anhui* 241000, *China*;

[3] *Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment, Ministry of Education, Wuhu, Anhui* 241000, *China*

**Abstract** Aiming at the problems of missing detection, false detection, and low detection accuracy of the Faster R-CNN target detection network, a soft non-maximum suppression (Soft-NMS) fusion attention mechanism and the Faster R-CNN (Faster Region-Convolutional Neural Network) target detection algorithm is proposed. In order to enhance the global important feature extraction and weaken the irrelevant feature in the feature map by the Faster R-CNN target detection algorithm, an attention mechanism is firstly introduced into the network. Second, aiming at the problem of local information loss caused by the bottleneck structure formed by two fully connected layers in the attention mechanism, a non-dimensional-reduction channel attention and spatial attention series module that can be trained end-to-end with the convolutional neural network is constructed. Then, a Soft-NMS is introduced to replace

**收稿日期**：2021-01-05；**修回日期**：2021-01-27；**录用日期**：2021-03-03

**基金项目**：安徽高校省级自然科学研究重点项目(KJ2019A0162)、安徽省自然科学基金(2108085MF197，1708085MF154)、检测技术与节能装置安徽省重点实验室开放基金资助项目(DTESD2020B02)

**通信作者**：*fswang@ahpu.edu.cn

the traditional non-maximal suppression (NMS) algorithm after the regional suggestion network, which can reduce the target missing detection and improve the location accuracy. Finally, the error detection rate is introduced into the evaluation criteria to further verify the performance of the model. Experimental results show that the Faster R-CNN algorithm based on ResNet-50 can effectively reduce the missed detection and false detection and improve the location accuracy, and the average detection accuracy is significantly improved.

**Key words** optics in computing; target detection; attention mechanism; NMS; CNN; Faster R-CNN

**OCIS codes** 200.4260; 150.1135; 110.2970; 100.4996

## 1 引 言

目标检测是计算机视觉中的一个重要研究方向，主要用于定位与识别图像和视频中的目标物体。近年来，目标检测已经在交通管控[1]、战场感知[2-3]和无人驾驶[4-5]等领域得到了广泛的应用。但是随着图像数据的数量越来越多，种类也越来越丰富，科研人员对目标检测的准确性和精度的要求越来越高，所以如何提升检测的准确性和精度具有重要意义。

目标检测算法可以分为基于传统的目标检测算法和基于深度学习的目标检测算法两类。其中，传统的目标检测算法适用于有明显特征和背景简单的情形，但是其在手工设计特征的过程中需要大量的先验知识，并且泛化能力不足，导致其通用性受到限制。基于深度学习的目标检测算法是从大量数据中自动学习特征，其泛化能力得到了显著提升，根据有无候选框生成阶段又可分为单阶段[6-7]和两阶段[8-11]目标检测算法，本文主要针对两阶段目标检测网络 Faster R-CNN(Faster Region-Convolutional Neural Network)[11]进行改进。对于两阶段目标检测而言，Girshick 等[8]提出了目标检测网络 R-CNN。与传统目标检测网络相比，R-CNN 目标检测网络中的 CNN[12-13]在 PASCAL VOC 数据集[14]上有更好的检测性能，但是 R-CNN 需要输入固定尺寸的图片，这会导致图片产生不必要的形变，造成检测效果差。2014 年，He 等[9]提出了 SPPNet(Spatial Pyramid Pooling Network)目标检测网络，该网络可以输入任意大小的图片。然而，与 R-CNN 相同，SPPNet 的训练步骤多，即训练 CNN 来提取特征，然后训练 SVM(Support Vector Machine)来分类这些特征，这就需要占据巨大的存储空间且处理速度较慢。针对 R-CNN 和 SPPNet 存在的问题，Girshick[10]提出了新的目标检测网络，其是基于 VGG-16(Visual Geometry Group-16)[12]来实现的，在训练、测试速度以及检测精度上都得到了显著的提高，但是 Fast R-CNN 目标检测网络采用选择性搜索算法来提取候选框，这会造成时间损失。Ren 等[11]提出 Faster R-CNN 目标检测网络，使用区域生成网络(RPN)代替 Fast R-CNN 的选择搜索算法来提取候选框，在速度和精度上都得到了显著的提升。

然而 Faster R-CNN 目标检测网络是通过卷积神经网络来提取特征以获得特征图，卷积层的设计保留了图片中的局部信息，但是卷积核固有的局部性使其无法得到图片中的全局特性，导致部分信息的丢失，造成精度的损失。此外，利用传统的非极大抑制(NMS)[15]来去除目标检测中的重复框，这会将与目标框相邻的检测框的分数强制归零，导致漏检。基于以上问题，为了提高 Faster R-CNN 目标检测网络的检测精度，本文对 Faster R-CNN 目标检测网络提出了如下改进。第一，在 Faster R-CNN 目标检测模型中引入基于现有注意力机制的改进卷积注意力网络，以获得图片更多的细节信息来提高模型的目标检测精度；第二，引入软非极大抑制(Soft-NMS)方式[16]来减少漏检并提高目标的定位精度。

## 2 相关工作

### 2.1 Faster R-CNN 模型

Faster R-CNN[11]目标检测模型如图 1 所示，主要包括 5 个步骤。1)首先将输入图片的长边和短边进行同比例缩放，并保证短边不超过 600 pixel，长边不超过 1000 pixel。2)将图片输入到主干特征提取网络中进行目标的特征提取以得到特征图，用于接下来的分类和回归预测。3)由于在主干网络中得到的特征图与输入图片存在一定的映射关系，特征图上的每一个像素对应着输入图像中一个区域的中心。将步骤 2)获得的特征图输入 RPN 中，利用锚框(anchor)机制在特征图上进行滑窗操作，在特征图中每个像素点对应原图的区域生成 9 个可能存在目标的候选框。然后利用分类分支来判断候选框是否包含目标；利用边界框回归分支来获得精确的建议框，在模型测试的过程中，使用NMS根据分类的

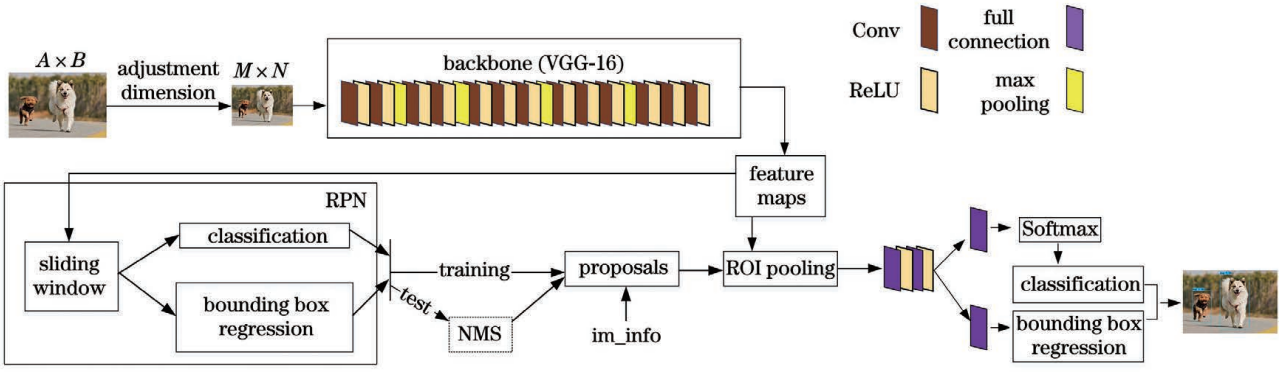


图 1　Faster R-CNN 目标检测模型

Fig. 1　Faster R-CNN target detection model

得分对建议框进行排序以去除重叠的建议框，输出得分较高的建议框并作为感兴趣区域(ROI)；在模型的训练过程中，使用 Smooth_L1_Loss 函数进行建议框位置的调整，以获得精确的建议框。在建议层，先利用尺度映射函数(im_info)将包含目标的 anchor 从 $M\times N$ 大小的尺度映射回 $A\times B$ 大小的原图中，用来判断包含目标的 anchor 是否大范围超过边界，剔除严重超出边界的 anchor，然后利用边界框回归来修正得到最终的建议框。4)将步骤 3)提取的建议框所对应的特征图划分为均等的 7×7 块，然后对每一块进行最大池化，可以将区域特征图大小固定为 7×7。5)输入步骤 4)得到的特征图，通过全连接层与 Softmax 分类器来计算建议框包含目标的概率；同时利用边界框回归对包含目标的建议框进行微调，以输出最终结果。

### 2.2　注意力机制

注意力机制是增强卷积神经网络性能的一种重要方式。SE-Net (Squeeze-and-Excitation Networks)[17]中设计了通道注意力的有效机制，有效建立了特征之间的空间相关性，并提升了卷积神经网络的性能。随后，基于 SE-Net 模型提出了卷积注意力机制(CBAM)[18]。与 SE-Net 不同，CBAM 中包含通道注意力和空间注意力两个模块。首先将输入的特征图通过一个通道注意力模块来得到加权结果，再经过空间注意力模块的加权得到最终的特征图。同样基于 SE-Net 模型，Wang 等[19]提出了 ECA-Net (Efficient Channel Attention Network)，其采用一维卷积层代替 SE-Net 的两个全连接层所组成的瓶颈结构，避免因降维造成细节信息的损失，并且改进的 ECA-Net 有效提升了卷积神经网络的性能。

### 2.3　NMS 算法

NMS 是计算机视觉中许多检测算法的重要组成部分。在目标检测任务中，NMS 主要就是通过迭代的形式不断将最大得分的框与其他框进行交并比(IoU)操作，并去除 IoU 值较大的框，具体可以分为以下 7 步。

1) 将所有的框按类别划分以去除背景类。

2) 对于每个目标类的边界框，按照分类置信度降序排列。

3) 在某一类中，选择置信度最高的边界框 $A'$ 并保留。

4) 逐一计算边界框 $A'$ 与剩余框 $B'$ 的 IoU，若 IoU 值大于阈值，则去除 $B'$。

5) 重复步骤 3)和步骤 4)，直到完成一个目标类的迭代。

6) 重复步骤 2)～5)，直到完成所有目标类的 NMS 处理。

7) 输出最终需要的框，算法结束。

## 3　本文算法

### 3.1　注意力模块的改进

综合 2.2 节所述的三个注意力机制，本文利用 ECA-Net 有效提高卷积神经网络性能的特点，对 CBAM 中的通道注意力机制进行一个非降维的操作，然后将改进的 CBAM 串接在 ResNet-50 (Residual Network-50)的最后一个 Identity block 之后进行前向传播。

对于 CBAM 的通道注意力模块，输入的特征图首先分别经过全局最大值池化和全局平均池化来聚合特征映射的空间信息，然后经过由两个全连接层组成的瓶颈结构来建模通道间的相关性，最后经过 Sigmoid 非线性激活函数来产生每一通道的权重，这样可以有效降低维度，但是降维操作会造成局部信息的丢失，并且远距离获取的信息没有相关性而

导致网格效应[20]，造成精度下降。基于以上分析，本文使用一维卷积层代替两个全连接层来完成权重的计算以避免降维，在减少维度损失的同时保证了通道之间的相关性。改进后的卷积注意力结构在网络中的应用方式如图 2 所示，其中 $M_c \in \mathbb{R}^{C\times1\times1}$ 表示基于空间压缩的通道注意力模块，$M_s \in \mathbb{R}^{1\times H\times W}$ 表示基于通道压缩的空间注意力模块，$C$、$H$ 和 $W$ 分别表示特征图的通道数、高和宽，⊗表示逐元素乘法，$\sigma$ 表示 Sigmoid 非线性激活函数，$C_{1D}$ 为一维卷积，⊕为相加操作。

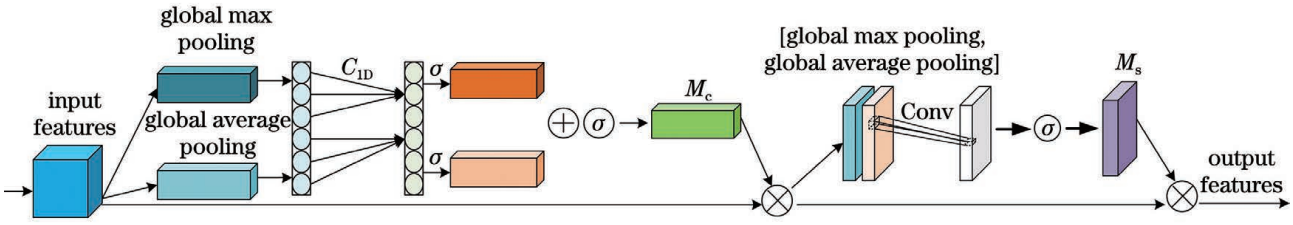


图 2　改进的卷积注意力机制的结构

Fig. 2　Structure of improved convolutional attention mechanism

改进的卷积注意力机制具体可表示为

$$F' = M_s[M_c(F) \otimes F] \otimes [M_c(F) \otimes F], \tag{1}$$

式中：$F \in \mathbb{R}^{C\times H\times W}$ 表示注意力机制模块的输入特征图；$F' \in \mathbb{R}^{C\times H\times W}$ 表示最终输出。

对于通道注意力模块 $M_c \in \mathbb{R}^{C\times1\times1}$，分别经过基于宽和高的全局最大池化和全局平均池化可以产生两个不同空间的上下文描述，记为 $F^c_{avg}$ 和 $F^c_{max}$；然后分别通过卷积核大小 $k=9$ 的一维卷积来计算权重，记为 $C_{1Dk}$；最后将权重加和合并，并使用 Sigmoid 非线性激活函数来输出最终的通道注意力特征图，数学表达式为

$$M_c(A) = \sigma\{C_{1Dk}[F^c_{avg}(A)] + C_{1Dk}[F^c_{max}(A)]\}。 \tag{2}$$

对于空间注意力模块 $M_s \in \mathbb{R}^{1\times H\times W}$，第一步将经过通道注意力模块产生的特征图作为本模块的输入特征图，并使用两个池化操作对通道信息进行压缩，得到的结果分别记为 $F^s_{avg} \in \mathbb{R}^{1\times H\times W}$ 和 $F^s_{max} \in \mathbb{R}^{1\times H\times W}$；第二步将 $F^s_{avg}$ 和 $F^s_{max}$ 合并可以得到一个双通道的特征图；第三步利用 7×7 大小的标准卷积核 $f^{7\times7}$ 将双通道的特征图降维成单通道，目的是保持前后维度一致，便于特征融合；第四步经过 Sigmoid 函数输出最终的空间注意力特征权值，可简单描述为

$$M_s(B) = \sigma[f^{7\times7}(F^s_{avg}; F^s_{max})]。 \tag{3}$$

### 3.2　Soft-NMS 算法

根据 2.3 节 NMS 算法中的步骤 4)可以知道，当 $B'$ 与置信度最高的 $A'$ 相交时，并且两边框的 IoU 值在预设的重叠阈值 $N_t$ 之内，则将 $B'$ 强制删除，这样可能会导致检测不到 $B'$，从而造成漏检。因此，本文在 Faster R-CNN 目标检测网络中引入 Soft-NMS 算法来代替 NMS 算法。

传统的 NMS 处理方式可以通过以下的分数重置函数来直观的表示，即

$$s_i = \begin{cases} s_i, & I_{IOU}(A', B'_i) < N_t \\ 0, & I_{IOU}(A', B'_i) \geqslant N_t \end{cases}, \tag{4}$$

式中：$s$ 为置信度得分；$i$ 为除得分最大的 $A'$ 框以外，剩余框以得分从高到底的排序的序号。与 NMS 算法相比，Soft-NMS 算法会对 $B'$ 的检测分数进行衰减而非彻底移除。文献[16]首先提出了线性的分数重置函数来解决漏检，对传统的 NMS 分数重置函数进行如下改进

$$s_i = \begin{cases} s_i, I_{IOU}(A', B'_i) < N_t \\ s_i[1 - I_{IOU}(A', B'_i)], I_{IOU}(A', B'_i) \geqslant N_t \end{cases}。 \tag{5}$$

通过上述函数，将大于阈值 $N_t$ 的检测分数衰减作为关于 $A'$ 重叠度的线性函数，因此当框 $B'$ 远离框 $A'$ 时，$B'$ 不会受到影响。但是线性的分数重置函数不一定是一个连续的函数，当框 $A'$ 和框 $B'$ 的 IoU 值达到 $N_t$ 时，检测序列可能突然变化。此外，当 $A'$ 和 $B'$ 的重叠度较低时，检测分数应该逐渐递增。最后提出高斯分数重置函数来解决检测序列可能突然变化的情况，最终的改进如下

$$s_i = s_i \exp\left[-\frac{I_{IOU}(A', B'_i)^2}{\sigma}\right]。 \tag{6}$$

针对传统非极大抑制算法存在的问题，本文采用含高斯分数重置函数的 Soft-NMS 算法，对 Faster R-CNN 目标检测网络中的传统 NMS 算法进行替换。结合(6)式可知，当两个框的重合度越高时，$s_i$ 的取值会越小，即降低相应框的得分可以避免因强制删除相应框而造成漏检的情况，从而提高目标检测的精度。

### 3.3　改进后的 Faster R-CNN 模型

图 3 为所提出的改进的 Faster R-CNN 目标检测模型。在改进的模型中，其主干特征提取网络为 ResNet-50。为了可以使用迁移学习权重，在不改变主干特征提取网络 ResNet-50 结构的前提下，首先在最后一个 Identity block 之后直接利用前向传播将改进的注意力模块串联接入模型中，再进行模型训练；然后将传统非极大抑制 NMS 替换为 Soft-NMS 进行测试。

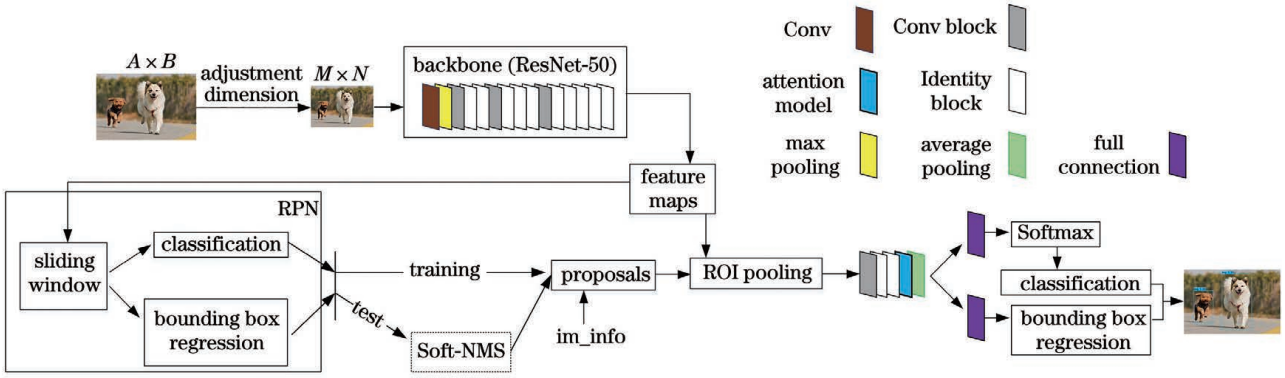


图 3　改进的 Faster R-CNN 模型

Fig. 3　Improved Faster R-CNN model

## 4　实验结果与分析

实验使用的操作系统为 Windows10，GPU 选用 Nvidia GeForce RTX 2080Ti(11 GB)，处理器为英特尔 Core i9-10900 @ 3.7 GHz，深度学习框架为 Pytorch 1.2.0。

### 4.1　数据集和评价标准

在实验中采用 PASCAL VOC 数据集进行训练和测试，利用 PASCAL VOC2007 和 PASCAL VOC2012 训练集中的 16551 张图片进行训练，利用 PASCAL VOC2007 测试集中的 4952 张图片进行测试，数据统计如表 1 所示。其中 PASCAL VOC2007 和 PASCAL VOC2012 数据集共有 4 个大类，如 vehicle、household、animal 和 person，总计 20 个小类。

表 1　VOC2007 和 VOC2012 数据集中训练和测试数据的统计

Table 1　Statistics of training and test data in VOC2007 and VOC2012 datasets

| Dataset | Training | | Test | |
|---|---|---|---|---|
| | Image | Object | Image | Object |
| VOC2007 | 5011 | 12608 | 4952 | 12032 |
| VOC2012 | 11540 | 27450 | 0 | 0 |
| Total | 16551 | 40058 | 4952 | 12032 |

为了检验改进的 Faster R-CNN 目标检测模型的性能，本文从客观评价和主观评价两个方面对模型的性能进行评估。对于客观评价结果，使用 AP(Average Precision)及 MR(Miss Rate)来对每一类检测结果进行评估，并利用 mAP(mean AP)，即所有目标类 AP 的平均值来衡量整个模型的性能；对于主观评价结果，可以通过对比改进和未改进网络处理后的图片来评估模型的性能，具体来说就是比较目标框的定位精确度和是否存在漏检和误检。

其中检测平均精度 AP 是 PR 曲线的面积，由精度($P$)和召回率($R$)组成，精度和召回率的计算公式为

$$P=\frac{x_{\mathrm{TP}}}{x_{\mathrm{TP}}+x_{\mathrm{FP}}},\tag{7}$$

$$R=\frac{x_{\mathrm{TP}}}{x_{\mathrm{TP}}+x_{\mathrm{FN}}},\tag{8}$$

式中：$x_{\mathrm{TP}}$ 表示正样本被正确识别为正样本的数目；$x_{\mathrm{FP}}$ 表示负样本被错误识别为正样本的数目；$x_{\mathrm{FN}}$ 表示正样本被错误识别为负样本的数目。对于误检率 MR，先计算每一幅图像的误检率 $x$，然后计算每一类别的误检率 MR，表达式为

$$x=\frac{x_{\mathrm{FP}}}{x_{\mathrm{FP}}+x_{\mathrm{TN}}},\tag{9}$$

$$M_{\mathrm{MR}}=2^{\left[\lg(x_1)+\lg(x_2)+\cdots+\lg(x_n)\right]\div n},\tag{10}$$

式中：$x_{\mathrm{TN}}$ 表示负样本被正确识别为负样本的数目。

### 4.2　实验过程

为了验证改进的注意力机制的有效性，实验中分别在 SSD[7]、YOLOv4 (You Only Look Once v4)[21]和基于 ResNet-50 的 Faster R-CNN 目标检测算法中加入改进的注意力机制并进行对比。本文

改进的 Faster R-CNN 目标检测模型的主干特征提取网络采用 ResNet-50，实验中保留了基于 ResNet-50 的 Faster R-CNN 目标检测网络的参数设置，改进的模型在不改变主干特征提取网络结构的前提下，利用迁移学习的权重来训练模型。在分类回归之前嵌入注意力机制，并修改注意力机制的输入通道数为 2048。此外，使用 Soft-NMS 代替传统的非极大抑制算法进行测试可以减少漏检。

为了更加充分地验证本文改进的 Faster R-CNN 目标检测算法的有效性，实验模型的比较如表 2 所示，其中“√”表示模型中包含相应模块，“-”表示模型不包含相应模块。

表 2 实验模型结构的对比

Table 2 Comparison of experimental model structure

| Serial number | Model | Backbone | Channel attention | Spatial attention | Dimension reduction | NMS | Soft-NMS |
|---|---|---|---|---|---|---|---|
| 0 | FR[11] | VGG-16 | - | - | - | √ | - |
| 1 | FR | ResNet-50 | - | - | - | √ | - |
| 2 | +Soft | ResNet-50 | - | - | - | - | √ |
| 3 | +SE | ResNet-50 | √ | - | √ | √ | - |
| 4 | +ECA | ResNet-50 | √ | - | - | √ | - |
| 5 | +CBAM | ResNet-50 | √ | √ | √ | √ | - |
| 6 | Ours1 | ResNet-50 | √ | √ | - | √ | - |
| 7 | Ours2 | ResNet-50 | √ | √ | - | - | √ |

表 2 中，实验 1 是将 ResNet-50 作为主干特征提取网络的 Faster R-CNN 目标检测模型；实验 2 是在实验 1 的网络模型中引入了 Soft-NMS；实验 3～5 分别是基于实验 1 的网络模型引入不同的注意力模块，即 SE-Net、ECA-Net 和 CBAM；实验 6 是基于实验 1 引入了本文改进的注意力机制的网络模型；实验 7 为本文最终的网络模型，即同时引入 Soft-NMS 和本文改进的注意力机制。

### 4.3 客观评价结果

为了验证本文改进的注意力机制模块的有效性，实验中将本文改进的注意力机制分别嵌入到 SSD[7]、YOLOv4[21] 和基于 ResNet-50 的 Faster R-CNN 目标检测算法中，实验结果如表 3 所示，其中 FR、SSD 和 YOLOv4 为基于 Pytorch 框架复现的原算法，Ours1、SSD+ 和 YOLOv4+ 分别表示在相应原算法中引入本文改进的注意力机制的算法，Variation 为每组实验的变化情况，“+”表示改进算法相对于原算法的增长数，“-”表示改进算法相对于原算法的下降数。

表 3 改进的注意力机制有效性验证实验的结果

Table 3 Results of validation experiment of improved attention mechanism unit: %

| Category | FR | Ours1 | Variation | Category | SSD | SSD+ | Variation | Category | YOLOv4 | YOLOv4+ | Variation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cat | 87.2 | 89.2 | +2.0 | Cat | 89.7 | 89.6 | -0.1 | Cat | 90.3 | 90.2 | -0.1 |
| Car | 86.6 | 85.4 | -1.2 | Car | 84.9 | 85.1 | +0.2 | Car | 94.8 | 95.1 | +0.3 |
| Horse | 84.9 | 86.6 | +1.7 | Horse | 89.1 | 89.4 | +0.3 | Horse | 91.7 | 91.0 | -0.7 |
| Dog | 83.6 | 85.6 | +2.0 | Dog | 85.7 | 86.1 | +0.4 | Dog | 87.1 | 88.9 | +1.8 |
| Bus | 80.3 | 85.3 | +5.0 | Bus | 84.2 | 83.9 | -0.3 | Bus | 91.3 | 92.6 | +1.3 |
| Train | 82.5 | 82.5 | 0 | Train | 87.1 | 87.2 | +0.1 | Train | 93.1 | 92.6 | -0.5 |
| Motorbike | 83.5 | 81.0 | -2.5 | Motorbike | 84.2 | 84.5 | +0.3 | Motorbike | 92.1 | 92.1 | 0 |
| Bicycle | 79.6 | 80.9 | +1.3 | Bicycle | 86.7 | 87.1 | +0.4 | Bicycle | 90.7 | 91.1 | +0.5 |
| Person | 79.8 | 79.7 | -0.1 | Person | 81.4 | 81.2 | -0.2 | Person | 91.2 | 91.1 | -0.1 |
| Aeroplane | 78.1 | 77.6 | -0.5 | Aeroplane | 76.2 | 77.5 | +1.3 | Aeroplane | 87.9 | 90.4 | +2.5 |

表 3 续

| Category | FR | Ours1 | Variation | Category | SSD | SSD+ | Variation | Category | YOLOv4 | YOLOv4+ | Variation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sheep | 75.2 | 76.9 | +1.7 | Sheep | 75.3 | 77.1 | +1.8 | Sheep | 87.4 | 87.7 | +0.3 |
| Bird | 74.1 | 75.3 | +1.2 | Bird | 75.4 | 75.0 | −0.4 | Bird | 87.4 | 86.0 | −1.4 |
| Cow | 74.7 | 75.1 | +0.4 | Cow | 78.5 | 79.6 | +1.1 | Cow | 91.5 | 92.1 | +0.6 |
| Tvmonitor | 73.3 | 72.5 | −0.8 | Tvmonitor | 76.6 | 76.7 | +0.1 | Tvmonitor | 89.0 | 90.1 | +1.1 |
| Diningtable | 72.3 | 73.0 | +0.7 | Diningtable | 76.3 | 80.1 | +3.8 | Diningtable | 80.9 | 81.9 | +1.0 |
| Sofa | 70.4 | 75.2 | +4.8 | Sofa | 78.0 | 80.4 | +1.6 | Sofa | 77.1 | 79.2 | +2.1 |
| Boat | 65.7 | 66.4 | +1.3 | Boat | 67.3 | 66.4 | −0.9 | Boat | 75.4 | 79.0 | +3.6 |
| Chair | 54.2 | 53.4 | −0.8 | Chair | 59.7 | 61.2 | +0.5 | Chair | 73.2 | 71.9 | −1.3 |
| Bottle | 52.1 | 53.6 | +1.5 | Bottle | 49.2 | 48.8 | −0.4 | Bottle | 82.1 | 80.8 | −1.3 |
| Pottedplant | 46.0 | 44.4 | −1.6 | Pottedplant | 47.2 | 47.6 | +0.4 | Pottedplant | 60.0 | 60.6 | +0.6 |
| mAP | 74.2 | 75.0 | +0.8 | mAP | 76.6 | 77.2 | +0.6 | mAP | 85.7 | 86.2 | +0.5 |

从衡量模型整体性能的 mAP 值可以发现，Faster R-CNN 目标检测模型加入本文改进的注意力机制后，mAP 值提高了 0.8 个百分点；SSD 目标检测算法加入本文改进的注意力机制后，mAP 值提高了 0.6 个百分点；对于 YOLOv4，YOLOv4 加入本文改进的注意力机制后，mAP 得到了 0.5 个百分点的提升。最后，从单目标类的检测精度可以发现，在三个不同的目标检测模型中嵌入本文改进的注意力机制，均可以提升大部分目标类的检测精度，其中将改进的注意力机制引入 Faster R-CNN 中，可以使单目标类的检测精度得到了明显的提升，并且最高约得到了 5 个百分点的精度提升。综上可以发现，本文所提出的改进注意力机制分别应用在 Faster R-CNN、YOLOv4 和 SSD 中均能得到一定的精度增益，也充分说明本文所提改进的注意力机制模块具有有效性和鲁棒性。

为了验证本文改进的 Faster R-CNN 目标检测算法的有效性，分别对表 2 中每个网络模型进行实验，mAP 结果如表 4 所示，其中“～600”表示短边不超过 600 pixel。

实验中，首先将基于 ResNet-50 的 Faster R-CNN 目标检测网络分别引入 Soft-NMS、SE-Net、ECA-Net 和 CBAM 中进行训练和测试。其中引入 Soft-NMS 的目标检测网络相比于原网络，mAP 值提升了 0.3 个百分点，这在一定程度上提升了目标检测的精度，这是因为传统的非极大抑制方法在处理重叠度较高的两个边界框的过程中，强制删除置信度较低的框会导致物体漏检，从而造成精度下降，而 Soft-NMS 算法是降低相应框的置信度而不是删除，这种做法在一定程度上避免了目标漏检，并提高了目标检测的精度。在 Faster R-CNN 目标检测网络中引入三种不同的注意力机制发现，注意力机制可以有效提升目标检测的精度，这是因为注意力机制可以显式地建模特征通道以及空间之间的相互依赖关系，通过学习的方式来自动获取每个特征通道以及空间的重要程度，然后依照重要程度可以有效定位到特征图中感兴趣的信息，并抑制无用信息。但是，SE-Net 和 CBAM 均利用两个全连接层所组成的瓶颈结构来建立通道间的相关信息，有效降低了维度，但是降维的处理会造成细节信息的损失；此外，将 SE-Net、ECA-Net 与 CBAM 卷积注意力机制进行比较，SE-Net 和 ECA-Net 只关注了通道注意力机制，而忽略了空间注意力机制，这样会造成一定的信息损失。本文通过

表 4 实验模型在 VOC2007 测试集上的 mAP 对比

Table 4 MAP comparison of experimental models on VOC2007 test set

| Model | Backbone | Input image size / (pixel×pixel) | mAP /% |
|---|---|---|---|
| FR[11] | VGG-16 | ～600×1000 | 73.2 |
| FR | ResNet-50 | ～600×1000 | 74.2 |
| +Soft | ResNet-50 | ～600×1000 | 74.5 |
| +SE | ResNet-50 | ～600×1000 | 75.2 |
| +ECA | ResNet-50 | ～600×1000 | 74.8 |
| +CBAM | ResNet-50 | ～600×1000 | 74.8 |
| Ours1 | ResNet-50 | ～600×1000 | 75.0 |
| Ours2 | ResNet-50 | ～600×1000 | 75.9 |

对三种注意力机制的对比分析与改进，提出了一种避免降维的自适应通道、空间卷积注意力机制，并嵌入到 Faster R-CNN 目标检测模型中，结果如表 2 实验 6 所示。实验结果表明，该方法取得了比引入 ECA-Net 和 CBAM 更好的效果，mAP 值提高了 0.2 个百分点。

基于以上实验结果的对比分析，本文在网络中嵌入了改进的注意力机制，并引入 Soft-NMS 算法来提高网络性能。通过将本文算法的实验 7 与实验 1 对比发现，改进算法的 mAP 值增长了 1.7 个百分点，并记录每个实验的 20 个目标类的检测平均精度 AP，结果如表 5 所示。

表 5　20 个目标类的 AP 对比

Table 5　AP comparison of 20 target classes　　unit：%

| Category | FR[11] | FR | +Soft | +SE | +ECA | +CBAM | Ours1 | Ours2 |
|---|---|---|---|---|---|---|---|---|
| Cat | 86.4 | 87.2 | 87.4 | 89.9 | 88.6 | 87.4 | 89.2 | 90.1 |
| Car | 84.7 | 86.6 | 87.1 | 85.6 | 85.2 | 85.1 | 85.4 | 86.1 |
| Horse | 84.6 | 84.9 | 85.2 | 86.6 | 86.1 | 86.6 | 86.6 | 86.8 |
| Dog | 84.8 | 83.6 | 84.0 | 87.8 | 84.1 | 82.6 | 85.6 | 86.8 |
| Bus | 83.1 | 80.3 | 80.3 | 83.1 | 84.9 | 84.7 | 85.3 | 85.3 |
| Train | 83.0 | 82.5 | 84.2 | 84.1 | 81.9 | 83.0 | 82.5 | 84.4 |
| Motorbike | 77.5 | 83.5 | 82.9 | 82.5 | 83.4 | 80.8 | 81.0 | 82.5 |
| Bicycle | 79.0 | 79.6 | 80.4 | 81.5 | 83.3 | 81.5 | 80.9 | 82.6 |
| Person | 76.7 | 79.8 | 80.5 | 81.0 | 79.6 | 80.8 | 79.7 | 81.3 |
| Aeroplane | 76.5 | 78.1 | 78.9 | 75.9 | 78.4 | 76.9 | 77.6 | 78.7 |
| Sheep | 73.6 | 75.2 | 74.4 | 73.0 | 73.5 | 70.5 | 76.9 | 77.3 |
| Bird | 70.9 | 74.1 | 74.2 | 76.0 | 75.9 | 75.8 | 75.3 | 76.8 |
| Cow | 81.9 | 74.7 | 75.0 | 77.5 | 74.0 | 77.2 | 75.1 | 75.9 |
| Tvmonitor | 72.6 | 73.3 | 74.4 | 72.1 | 73.0 | 75.4 | 72.5 | 73.1 |
| Diningtable | 65.7 | 72.3 | 72.0 | 71.7 | 72.5 | 72.4 | 73.0 | 73.6 |
| Sofa | 73.9 | 70.4 | 70.9 | 73.5 | 71.1 | 71.6 | 75.2 | 76.2 |
| Boat | 65.5 | 65.7 | 65.9 | 66.1 | 66.4 | 66.1 | 66.4 | 66.7 |
| Chair | 52.0 | 54.2 | 54.1 | 53.6 | 53.7 | 54.1 | 53.4 | 54.1 |
| Bottle | 52.1 | 52.1 | 52.6 | 59.1 | 54.8 | 56.0 | 53.6 | 54.0 |
| Pottedplant | 38.8 | 46.0 | 46.0 | 44.2 | 45.9 | 45.6 | 44.4 | 46.1 |
| mAP | 73.2 | 74.2 | 74.5 | 75.2 | 74.8 | 74.8 | 75.0 | 75.9 |

表 5 是利用 PASCAL VOC2007 和 PASCAL VOC2012 的训练集进行联合训练，然后基于 PASCAL VOC2007 的测试集来评估 20 个目标类的检测平均精度。从表 5 可以看出，使用 Soft-NMS 替换传统的 NMS，相比于未改进的算法，提高了 70%的目标类的检测平均精度；重合度比较低的目标类的检测平均精度基本保持不变或者略有降低，如桌子、椅子和沙发等目标类；对于人、汽车和瓶子等在数据集中占重合度比较高的目标类，精度均得到了一定的提升；此外，分别嵌入 SE-Net、ECA-Net 和 CBAM 的目标检测网络，相比较于未改进的模型，精度分别得到了 60%、55%和 65%的提升。目标检测的精度会受到遮挡、目标过小和阴影等因素的影响，导致不同目标类的检测精度存在差距，但是引入注意力机制发现，对于小目标和遮挡等目标，检测精度均有提升，如猫、公共汽车、自行车和鸟的检测精度，分别最高提高了 2.7，4.6，3.7，1.9 个百分点；相比于引入现有的注意力机制，本文改进的注意力机制嵌入 Faster R-CNN 目标检测模型中得到了 60%的精度提升，其中对于边缘比较明显的目标

类，精度提升明显，如猫、马、狗、公共汽车和沙发分别得到了 2.0，1.7，2.0，5.0，4.8 个百分点的精度提升；对于边缘不明显的目标，检测精度的提升不明显或有所降低，如盆栽的检测精度降低了 1.6 个百分点。由表 5 可以看到，与其他算法相比，本文最终算法有效提高了 80%的目标类的检测平均精度，并拥有一定的优势。

为了进一步验证改进模型的性能，本文引入误检率，并将本文算法和原算法进行误检率 $R$ 比较，结果如图 4 所示，可以看出改进的算法明显降低了大部分类别的误检率，提高了总体的平均检测精度。

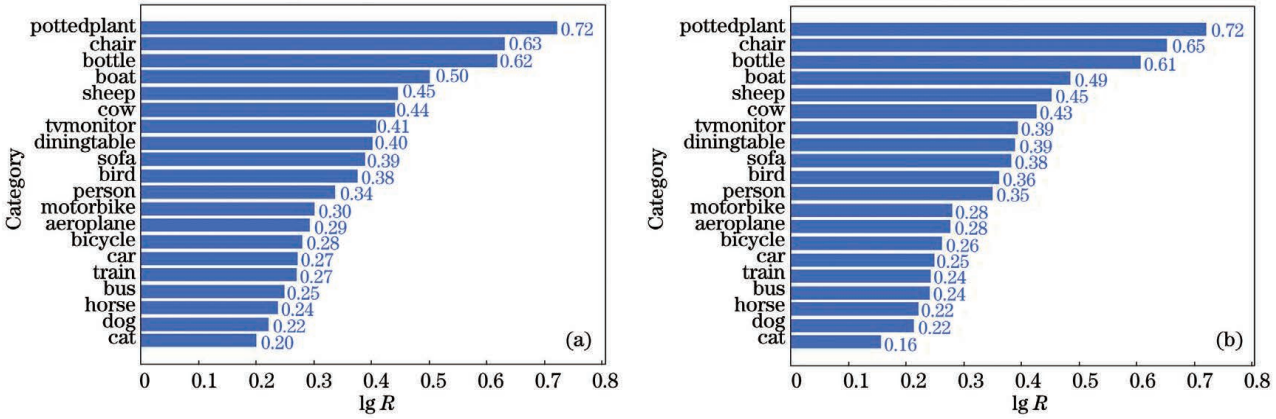


图 4　误检率比较。(a)原算法的误检率统计；(b)改进算法的误检率统计

Fig. 4　Comparison of false detection rate. (a) Error detection rate statistics of original algorithm; (b) error detection rate statistics of improved algorithm

为了验证本文算法的优越性，使用基于 VOC2007＋VOC2012 的混合数据集进行训练，并使用 VOC2007 测试集进行测试，将部分目标检测算法的 mAP 与本文算法进行比较，结果如表 6 所示，其中 DiCENet 指 Dimension-wise Convolutions for Efficient Networks。

表 6　不同算法的 mAP 比较

Table 6　mAP comparison of different algorithms

| Algotithm | Backbone | mAP /% |
|---|---|---|
| Ref. [11] | VGG-16 | 73.2 |
| Ref. [22] | ResNet-101 | 76.4 |
| Ref. [23] | ResNet-50 | 74.4 |
| Ref. [24] | ResNet-101 | 74.4 |
| YOLO[6] | Darknet | 63.4 |
| SSD300[7] | VGG-16 | 74.3 |
| Ref. [25] | DiCENet | 68.4 |
| Ref. [26] | ResNet-50 | 74.4 |
| Ours2 | ResNet-50 | 75.9 |

将表 6 不同算法所得到的 mAP 进行数据对比发现，本文提出的基于注意力机制和 Soft-NMS 的改进 Faster R-CNN 目标检测算法在检测精度上具有较好的表现，本文算法的精度相较于两阶段经典算法 Faster R-CNN[11] 有 2.7 个百分点的提升。表 6 中的文献[22-24]算法是基于文献[11]的改进算法，本文算法与文献[22-24]算法相比具有一定的竞争力，将本文算法与文献[6-7]算法进行比较，可以发现精度分别提升了 12.5 个百分点和 1.6 个百分点。此外，本文算法的检测精度比文献[25-26]算法分别提升了 6.5 个百分点和 1.5 个百分点。

### 4.4　主观评价结果

由于目标检测的任务是找出图像中所有感兴趣的目标，并确定目标的位置和大小。改进的算法和原算法在相同图像上所检测的结果会有所区别，为了更直观地体现改进的目标检测算法的性能，实验中随机抽取图片，先利用基于 ResNet-50 的 Faster R-CNN 目标检测模型即原算法进行检测，然后利用本文算法进行检测，实验检测结果的对比如图 5 所示。

通过观察图 5 中 image 1、image 3、image 5 和 image 9 的检测结果可以发现，原算法存在当两个物体重合度较高的漏检情况，而本文利用 Soft-NMS 算法有效减少了漏检。此外，针对原算法存在误检和定位不精确的问题，本文在目标检测网络中引入注意力机制来充分利用图片中有用的信息并抑制无关信息，从 image 4、image 8 和 image 9 可以发现本文算法降低了误检的情况，并且从image2和


图 5　图片检测比较。(a)原算法的检测结果;(b)改进算法的检测结果

Fig. 5　Image detection comparison. (a) Detection results of original algorithm; (b) detection results of improved algorithm

image 7 可以清楚看到本文算法有效提高了定位精度,但是改进的算法仍存在误检的情况,如将 image 5 中的牛检测成羊的问题以及 image 10 类似误检情况。综合来看,本文算法相对于原算法取得了较优的性能。

## 5　结　　论

本文提出了一种改进的 Faster R-CNN 目标检测网络,首先通过嵌入改进的卷积注意力机制可以使卷积神经网络输出特征图的全局特征;其次通过引入 Soft-NMS 有效降低因应用传统非极大抑制 NMS 算法而导致密集物体相邻框的漏检问题。本文算法与基于 ResNet-50 的 Faster R-CNN 目标检测算法相比,最终改进的目标检测网络在 VOC2007 的测试集上提高了 80%目标类的检测平均精度,其中最高得到了 5.8 个百分点的精度提升,此外最终改进的目标检测网络使 mAP 值提升了 1.7 个百分点,有效降低了漏检和误检概率,并且提高了感兴趣目标框的定位精度。

## 参　考　文　献

[1] Pan Q C, Zhang H H. Key algorithms of video target detection and recognition in intelligent transportation systems[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2020, 34(9): 2055016.

[2] Liu Y J, Yang F B, Hu P. Parallel FPN algorithm based on cascade R-CNN for object detection from UAV aerial images[J]. Laser & Optoelectronics Progress, 2020, 57(20): 201505.
刘英杰, 杨风暴, 胡鹏. 基于 Cascade R-CNN 的并行特征金字塔网络无人机航拍图像目标检测算法[J]. 激光与光电子学进展, 2020, 57(20): 201505.

[3] Cao Y J, Xu G M, Shi G C. Low altitude armored target detection based on rotation invariant Faster R-CNN[J]. Laser & Optoelectronics Progress, 2018, 55(10): 101501.
曹宇剑, 徐国明, 史国川. 基于旋转不变 Faster R-

CNN 的低空装甲目标检测[J]. 激光与光电子学进展, 2018, 55(10): 101501.

[4] Huang G, Liu X L. Automatic road marking extraction and classification method based on deep learning[J]. Chinese Journal of Lasers, 2019, 46(8): 0804002.
黄刚, 刘先林. 基于深度学习的道路标线自动提取与分类方法[J]. 中国激光, 2019, 46(8): 0804002.

[5] Wang K J, Zhao Y D, Xing X L. Deep learning in driverless vehicles[J]. CAAI Transactions on Intelligent Systems, 2018, 13(1): 55-69.
王科俊, 赵彦东, 邢向磊. 深度学习在无人驾驶汽车领域应用的研究进展[J]. 智能系统学报, 2018, 13(1): 55-69.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[7] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[8] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[9] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8691: 346-361.

[10] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[11] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations (ICLR), May 7-9, 2015, San Diego, CA, USA. [S.l.: s.n.], 2015: 1150-1210.

[13] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[14] Everingham M, Eslami S M A, Gool L, et al. The pascal visual object classes challenge: a retrospective [J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[15] Neubeck A, Van Gool L. Efficient non-maximum suppression[C]//18th International Conference on Pattern Recognition (ICPR'06), August 20-24, 2006, Hong Kong, China. New York: IEEE Press, 2006: 850-855.

[16] Bodla N, Singh B, Chellappa R, et al. Soft-NMS: improving object detection with one line of code[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 5562-5570.

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[18] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[19] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11531-11539.

[20] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions [C]// Proceeding of the 4th International Conference on Learning Representations (ICLR), May 2-4, 2016, San Juan, Puerto Rico. [S.l.: s.n.], 2016: 23-24.

[21] Alexey B, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2021-01-01]. https://arxiv.org/abs/2004.10934.

[22] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[23] Gao S H, Cheng M M, Zhao K, et al. Res2Net: a new multi-scale backbone architecture [J]. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662.

[24] Wang T, Yuan L, Zhang X P, et al. Distilling object detectors with fine-grained feature imitation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4928-4937.

[25] Mehta S, Hajishirzi H, Rastegari M. DiCENet: dimension-wise convolutions for efficient networks [EB/OL]. (2019-06-08)[2021-01-01]. https://arxiv.org/abs/1906.03516v3

[26] Shlok M, Anshul S, Ankan B, et al. Learning visual representations for transfer learning by suppressing texture [EB/OL]. (2020-11-03)[2021-01-01]. https://arxiv.org/abs/2011.01901.